VPTMdb: a viral post-translational modification database

Yujia Xiang[1,3], QuanZou[2*] and Lilin Zhao[1,4*]

[1] State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China; [2] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China; [3] University of Chinese Academy of Sciences, Beijing, China; [4] CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing, China

*To whom correspondence should be addressed.

Abstract

Motivation: In viruses, post-translational modifications (PTMs) are essential for their life cycle. Recognizing viral PTMs is very important for better understanding the mechanism of viral infections and finding potential drug targets. However, a database of comprehensive PTM information on viruses is not available.

Results: To fill this gap, we developed a viral post-translational modification database (VPTMdb) for collecting systematic information of viral PTM data. The VPTMdb contains 912 PTM sites that integrate 414 experimental-confirmed PTM sites with 98 proteins in 45 human viruses manually extracted from 162 publications and 498 PTMs extracted from UniProtKB/Swiss-Prot. Moreover, we created a novel sequence-based classifier named VPTMpre to help users predict viral protein phosphorylated sites. Finally, an online web server was implemented for users to download viral PTM data and predict phosphorylated sites of interest.

Availability and Implementation: VPTMdb can be accessed at http://vptmdb.com:8787/VPTMdb/

Contact: zhaoll@ioz.ac.cn or zouquan@nclab.net

1 Introduction

Post-translational modifications (PTMs) play a critical role in current proteomics research and regulate protein functions by altering protein interactions, stability, activity, and subcellular localization. Post-translation modifications of viral proteins are relevant throughout various stages of the pathogen life cycle, especially viral infections and genome replication. For example, during entry, the influenza virus carries unanchored ubiquitin chains to engage the host cell's aggresome system (Banerjee, et al., 2014). Once inside the host cell, viral PTMs regulating infecting process of HSV-1 encode ICP0 protein to degrade host proteins via ubiquitination and sumoylation (Randow and Lehner, 2009). In the viral life circle, the HIV-1 Tat protein ser-16 phosphorylated site regulates HIV-1 transcription (Ivanov, et al., 2018).

Therefore, knowledge of viral PTMs is of great significant to understanding the molecular mechanisms underlying viral infections and recognizing potential drug targets. In recent years, several studies have identified multiple viral PTMs (Kulej, et al., 2017; Scaturro, et al., 2018; Zheng, et al., 2018); thus, establishing a database to provide relevant knowledge is important.

However, few databases have been developed for systematically archiving and easily accessing the PTM sites data of viruses. VirPTM (Schwartz and Church, 2010) stores viral phosphorylated sites. ViralPhos (Huang, et al., 2013) is a support vector machine based predictor and database that provides outdated viral phosphorylated sites. To date, no databases have collected comprehensive PTM data of viral proteins.

To bridge the existing knowledge gap, we have built a viral post-translational modification database (VPTMdb) that first provides comprehensive experimentally verified viral PTM site data, including phosphorylation, sumoylation, glycosylation, acetylation, methylation, ubiquitination, neddylation and palmitoylation, and it includes 162 studies that have been manually viewed to extract PTM sites. In total, 912 PTM sites from 45 human viruses were obtained, which include 414 manual checked sites from PubMed as well as 498 sites from UniProtKB/Swiss-Prot.

In addition, to determine the functional association of viral PTM sites located in the human-virus protein-protein interaction (PPI) domains, human-virus PPI data were downloaded from the HPIDB 3.0 (Ammari, et al., 2016) and integrated into the VPTMdb database.

Moreover, PTM was predicted in other species with machine learning approaches (He, et al., 2018; Huang and Li, 2018). For viral protein serine modification sites identification, we implemented a novel feature-based classifier named VPTMpre into the VPTMdb to provide users with the ability to find viral protein phosphorylated sites. First, we compared several feature extraction methods using support vector machine via a 5-fold cross-validation method to obtain the best feature representative strategy. Second, for feature selection, we separately input the features extracted from the previous step into three machine learning predictors (support vector machine, random forest and naïve Bayes). Using the minimum redundancy and maximum relevance (mRMR) algorithm (Hanchuan, et al., 2005), the predictive performance of these three classifiers on different feature dimensions was compared. Subsequently, the features that performed well in all three classifiers were selected as the most meaningful and significant features and 68-dimensional features were obtained. The results of independent testing showed that 68D features performed

better. Overall, VPTMpre represents a powerful tool to predict viral protein phosphorylated sites.

The online web server is available at http://vptmdb.com:8787/VPTMdb/, and users can browse and download viral PTM data freely. Support vector machine, random forest and naïve Bayes were integrated into VPTMpre, and users are able to choose one machine learning model to predict possible phosphorylated sites of interest.

2 Methods

2.1 Data collection

There are three major steps in data collecting and pre-processing, which are described below.

First, we queried PubMed using the keyword search terms: (virus name) and (eight modification types) for studies published before Jan 01, 2020. As a result, 6052 papers were obtained, each of which was manually retrieved using the following standards: (i) the viral post-translational modifications were experimentally verified; and (ii) if two references contained the same PTM site, the earliest published study was retained. In total, 45 viruses, 162 papers and 414 PTMs were obtained.

Subsequently, 498 viral PTM data points from UniProtKB/Swiss-Prot were integrated into VPTMdb. For experimentally validated virus PTM types, the sites were extracted manually from the articles mentioned above. The protein sequences, UniProt ID and PMID were mainly extracted from NCBI, UniProt and PubMed. Finally, human-virus protein-protein interactions were collected from the HPIDB 3.0 based on viral strains in the VPTMdb.

2.2 Overview of viral phosphorylation sites prediction

Identifying viral protein PTM sites by experimental methods is still expensive and time consuming. Thus, predicting them in *silico* using bioinformatics approaches is necessary. To this end, a sequence-based classifier named VPTMpre was created to predict viral post-translational modification serine sites. Because threonine and tyrosine data are too few to train the model, we only predicted serine sites in this study.

Five main procedures were performed to build the VPTMpre predictor. (i) a balanced benchmark dataset was constructed using the Synthetic Minority Oversampling Technique (SMOTE) (Chawla, et al., 2002) sampling method (**Supplementary Material Si1**); (ii) various feature representative methods were compared to obtain an effective feature representation strategy, with support vector machine used as the base classifier in a 5-fold cross-validation approach to find the best feature groups; (iii) the predictive performance of three classifiers (svm, random forest, naïve Bayes) on different feature dimensions was compared using the Minimum redundancy and maximum relevance (mRMR) method, and the features that performed well in all three classifiers were selected as the most meaningful and significant features; (iv) a 10-fold random independent test was performed to evaluate the predictive performance of the three different classifiers (svm, random forest, naïve Bayes); and (v) VPTMpre was implemented in the online web server.

2.3 Data preparation and processing

All viral phosphorylation experimentally verified serine sites in our database were used as positive samples, and those not marked by any phosphorylation information on the same protein were considered negative samples. As a result, we obtained 182 phosphorylated serine residues as well as 2148 non-phosphorylated residues. Phosphorylated sites from UniProtKB/Swiss-Prot were regarded as the independent dataset, and they included 93 positive serine sites and 1878 negative serine sites. After using CD-HIT (identity set to 0.8) (Fu, et al., 2012) to remove redundant sequences, we obtained 129 positive sites and 1611 negative sites. The independent dataset contained 52 positive sites and 1072 negative sites (**Table 1**). These sequences were truncated to a 23-residue symmetrical window (-11 to 11) (**Supplementary Data Table**).

In order to eliminate the prediction bias caused by data imbalance, we re-sampled the training data by SMOTE methods and obtained 260 positive sites and 260 negative sites, which consisted of the training dataset. The negative test set from UniProtKB/Swiss-Prot was randomly divided into twenty parts. We randomly select ten negative subsets from the twenty parts and combined them with ten replicate positive sets to constitute ten independent test datasets (**Supplementary Materials Table S1**).

Table 1. Summary of training and independent datasets

Datasets	*Types*	*Total number*	*After deletion*	*After balanced*
Training set	Positive	182	129	260
	Negative	2148	1611	260
Independent set	Positive	93	52	52
	Negative	1878	1072	1072

2.4 Feature representation

To achieve a better classification effect, a key step is feature extraction, which means that a protein sequence is encoded as a numeric vector for machine learning model.

Amino acid composition (AAC). AAC is the frequency of 20 amino acids for a given sequence (Liu, 2019). This descriptor can be denoted as follows:

$$AAC = (A1, A2, A3, \ldots, A20) \quad (1)$$

where

$$Ai = \frac{Ri}{L} (i = 1,2,3, \ldots ,20) \quad (2)$$

Ri is the observed number of types i amino acid in a protein sequence. L is the length of protein. Thus 20 features were obtained, and sum of which is 1.

Binary profile. The binary profile transformed each amino acid into a 20-dimensional binary numerical vector. For instance, the alanine ('A') is deciphered as 10000000000000000000, cysteine ('C') is deciphered as 01000000000000000000, etc. Consequently, we obtained a 460-dimensional vector for this binary profile feature.

Conjoint triad (CTriad). The conjoint triad feature is sequence information for proteins. Twenty amino acid types are clustered into seven classes to construct the C-triad feature.

$$group1 = \{Ala, Cly, Val\}, group2 = \{\text{Ile, Leu, Phe, Pro}\} \quad (3)$$
$$group3 = \{Tyr, Met, Thr, Ser\},$$
$$group4 = \{His, Asn, Gln, Trp\}$$
$$group5 = \{Arg, Lys\}, group6 = \{Asp, Glu\},$$
$$group7 = \{Cys\}$$

First, protein sequences are encoded into a numerical vector using the AA groups list above. Subsequently, any three continuous AAs are regarded as a unit, and scanning along the sequences and counting the frequencies of each triad type is performed to obtain a 343-dimensional numerical vector.

For example, a protein sequence S contains L AA residues, which are expressed as follows:

$$S = A_1 A_2 A_3 A_4 A_5 \ldots A_L. \quad (4)$$

Then, we scan along the sequence with a slide window in three continuous residues:

$$A_1A_2A_3, A_2A_3A_4, A_3A_4A_5, A_4A_5A_6, \ldots, A_{L-2}A_{L-1}A_L \quad (5)$$

Finally, the C-triad feature of a protein is defined as the frequency of the corresponding triad type in that protein:

$$Ctriad = [f_1, f_2, f_3, f_4, \ldots, f_{343}]^T \quad (6)$$

where,

$$f_i = \frac{n_i}{L-2} \quad (7)$$

n_i is the occurrence number of the i-th triad type (i= 1, 2, ..., 343).

More detailed information about C-triad can be found in (Shen, et al., 2007).

Composition-Transition-Distribution (CTD). CTD clusters 20 amino acids into three groups: hydrophobic, neutral and polar. The CTD composition (CTD-C) calculates the composition values of hydrophobic, neutral and polar groups for a given sequence. The CTD transition (CTD-T) represents the percentage frequency of an amino acid of one particular property followed by an amino acid of another property. The CTD distribution (CTD-D) represents the distribution of each property for a given sequence. Each property has five distribution descriptors, which are the first residue, 25% residues, 50% residues, 75% residues, and 100% residues in the whole sequence of a given specific property. In this research, CTD-C, CTD-T, and CTD-D were used to encoded protein sequences and yielded 39, 39, and 195 features, respectively. More detailed information about CTD can be found in the literature (Govindan and Nair, 2011).

Enhanced grouped amino acid composition (EGAAC). EGAAC was first proposed by Chen et al. (Chen, et al., 2018) and is the improved version of GAAC features. GAAC divides 20 standard amino acids into five groups based on their physical and chemical properties. The formulation of GAAC is as follows:

$$f(g) = \frac{N(g)}{L}, g \in \{g1, g2, g3, g4, g5\} \quad (8)$$

$$N(g_i) = \sum N_i, i \in g \quad (9)$$

$$g1 = \{GAVLMI\}, g2 = \{FYW\}, g3 = \{KRH\}, g4 = \{DE\}, g5 = \{STCPNQ\} \quad (10)$$

where L is the length of sequence, $N(g)$ is the number of amino acids in group g, and N_i is the occurrence number of i-th amino acid type.

EGAAC scans along the sequence and calculates the GAAC values in a fixed-size window:

$$F(g) = \frac{N(g, win)}{N(win)}, g \in \{g1, g2, g3, g4, g5\} \quad (11)$$

where $N(g,win)$ is the number of amino acids in group g within a fixed-size window *win* and $N(win)$ is the window size. *win* ranges from 1 to 17. In this study, the window size was set to 5, and we finally obtained a 95-dimensional vector.

Z-Scale (ZSCALE). Z-scale is a feature descriptor that describes AAs' physicochemical properties. It was first published by Hellberg (Hellberg, et al., 1987), who introduced three z-scales (z1-z3), and then Sandberg et al. (Sandberg, et al., 1998) improved the original z-scale features by adding two more z-scale values, using 26 properties of 87 AAs. In this study, we employed the z-scale using five scales(z1-z5). The five z-scales are based on lipophilicity (z1), bulk (z2), polarity/charge (z3), electronegativity and heat of formation(z4), electrophilicity and hardness(z5), yielding a 115-dimensional numerical vector.

2.5 Feature selection and optimization

Generally, high-dimension biological features may be noisy, which led to poor prediction performance. However, feature selection is a good strategy to overcome feature redundancy. Feature selection means using a reduction algorithm to select the major features that are able to improve the performance of specific classifiers.

In this work, six descriptors and their combined features' performance were compared using 5-fold cross validation in the training data with the Support Vector Machine (SVM) method. Subsequently, the Minimum redundancy and maximum relevance (mRMR) method was chosen to select the most meaningful features. To investigate the predictive performance of three classifiers, we compared the different dimensions of features in the svm, random forest, naïve Bayes methods. The features that performed well in all three classifiers were selected as the most meaningful and significant features. The T-distributed Stochastic Neighbour Embedding algorithm was used to visualize the features(van der Maaten and Hinton, 2008).

2.6 Performance evaluation

Sensitivity (Sn), Specificity (Sp), F1-score, and Mathews Correlation Coefficient (MCC) were applied to estimate the prediction performance (**Supplementary Materials Si2**). In addition, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were used to evaluate the overall performance of the model. The ROC curve is a continuous line plotted by the false positive rate (FPR) as the X-coordinate and true positive rate (TPR) as the Y-coordinate. The higher the AUC value, the better the performance of the classifier.

2.7 Website implementation

The VPTMdb web interface was written in the R programming language using the Rshiny web development framework (Chang, et al., 2018). The MySQL database management system was used to store structured PTM data. The base machine learning predictor (such as SVM) was supported by the caret R package (Kuhn, 2020); the ROC curve was analysed using ROCR (Sing, et al., 2005); and MRMR and t-SNE were analysed using mRMRe (De Jay, et al., 2013) and Rtsne (Krijthe, 2015). Software ggplot2 was used to plot beautiful pictures (Wickham, 2016). The website is free and can be browsed in most modern browsers.

3 Results

3.1 Database contents

Figure 1A shows that the VPTMdb webserver consists of two parts: VPTM database and VPTMpre. The VPTM database currently includes 414 unique experimentally determined PTM sites with 8 modification types from 45 viruses. In summary, 162 manually checked references were collected in the database. Each entry in VPTMdb includes the (i) virus name, (ii) virus protein name in the UniProt database, (iii) viral modification site, (iv) residue sequences, and (v) PubMed id. PTM data from UniProtKB/Swiss-Prot contain two types: 199 phosphorylation sites and 299 glycosylation sites (N-lined and O-lined).

The research on post-translational modification of viruses has been increasing over the past decade according to the 6052 literatures we downloaded from PubMed. The greatest number of publications was related to viral phosphorylation, followed by glycosylation and methylation (**Supplementary Fig.S1**).

The statistics of experimentally verified sites in VPTMdb show that among eight PTM types, phosphorylated sites account for the most (484 sites, including 285 manually checked and 199 sites from UniProtKB/Swiss-Prot) at more than 50% of the total database. The top five viruses in the number of manually checked modification sites are HAdV-2 (51 phosphorylation sites), EBOV (29 phosphorylation, 1 sumoylation, 2 ubiquitination, 8 acetylation sites), HIV-1 (21 phosphorylation, 4 sumoylation, 2 ubiquitination, 5 acetylation and 3 glycosylation sites), H1N1 (19 phosphorylation, 3 sumoylation, 2 ubiquitination, 6 acetylation and 2 glycosylation sites), and HCV (10 phosphorylation, 1 sumoylation, 1 ubiquitination, 1 methylation 4 palmitoylation and 14 glycosylation sites) (**Fig. 1B**).

Human-virus PPI data were included in the VPTMdb, which are helpful to determine the potential function of PTMs during viral infections. PPI data in the VPTMdb contain 114 interactions with 93 proteins in 17 viruses. In the database first version, there are six post-translational modified proteins in PPI data. **Figure 1C** shows the distribution of six modified proteins in the protein-protein interaction network.

The web server involves five easy-to-use main pages: 'Home', 'Browse', 'Prediction', 'Download', and 'Help'. Each of these pages enables users to search, browse, predict and download data without any prerequisite knowledge. In the 'Browse' section, users can search the PTM data conveniently by typing keywords in the search box and download data freely, what is more, virus-human protein-protein interaction data are provided and visualized. The 'Prediction' page provides VPTMpre, a sequence-based machine learning predictor for phosphorylation serine site prediction. All data about virus PTM are stored in the 'Download' page for batched downloading. The 'Help' page contains a detailed tutorial to help users learn about VPTMdb.

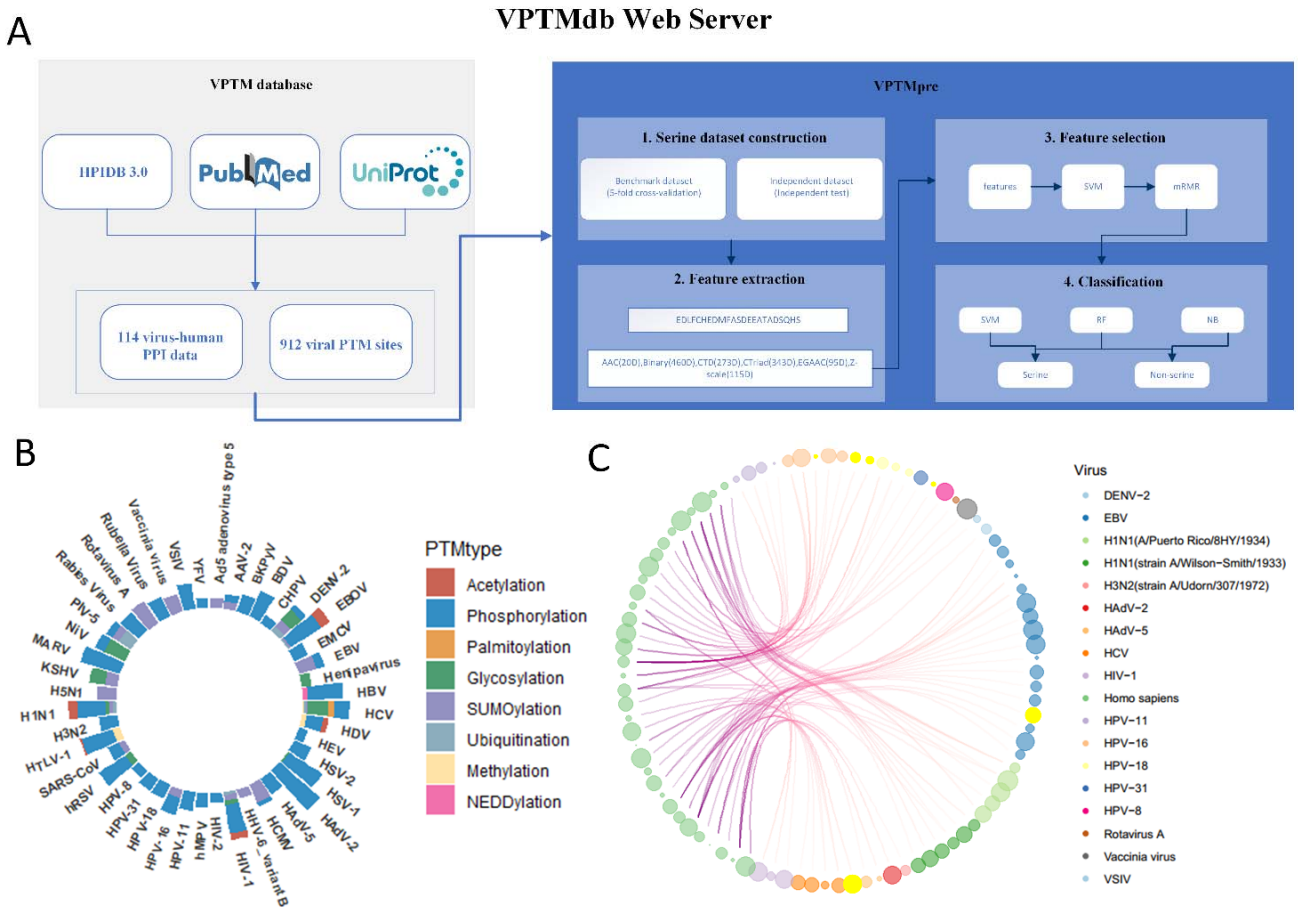

Figure 1. Overview of VPTMdb: (A) Framework of VPTMdb web server construction. First, PTM data were collected from PubMed and UniProt/Swiss-Prot. Then, VPTMpre was constructed to predict viral protein phosphorylated sites. (B) Statistics of viral PTM data in VPTMdb. (C) The virus-human protein-protein interaction network. Each node represents viral protein or human protein. Each edge represents virus-human association. The post-translational modified proteins in the network are highlighted with colour light-yellow.

3.2 Features comparison

The data in **Table 2** show that six features as well as their combinations were evaluated in SVM with a 5-fold cross-validation. AUC, F1-score and MCC were used as the performance evaluation indicators. The results declare that the z-scale, which captures the physical-chemical information of amino acids, is the best among the six single features (AUC=0.957, F1-Score=0.887, MCC=0.810). For BINARY, EGAAC and CTriad, their AUC values also achieved above 90.00%. Moreover, when we fused the features, the result showed that ZSCALE combined

with AAC features improved the sensitivity, F1-score and AUC by 8.40%, 1.5%, 0.1% compared with individual z-scale features.

However, the combination of EGAAC, BINARY, ZSCALE and CTriad features did not significantly enhance the model's performance, which suggests that high-dimensional features may include useless features that weaken the model performance. Among all the features, considering the three evaluation values of F1-score, MCC, AUC and dimensions, the AAC combined with the ZSCALE performed best, and the sensitivity, AUC and F1-score were higher than the single z-scale features. The independent test also shows that AAC combined with ZSCALE features significantly increased the AUC, F1-score, MCC, and Sn by 0.90%, 21.7%, 2.60%, and 25.0%, respectively (**Supplementary Table S2**).

Now, it is important to answer two questions: (i) what is the difference between phosphorylated sites and non-phosphorylated sites and (ii) which features contribute most to the viral phosphorylation protein? To this end, we analysed the z-scale feature information between phosphorylated sites and non-phosphorylated sites. Then, we selected the most important features from the combined features with the mRMR method and using svm, random forest and naïve Bayes to perform a predictive evaluation.

Table 2. Comparison of performance between the single features and fused features with the mRMR method

Features	Dim	Sn	Sp	MCC	F1	AUC
1.AAC	20	0.738	0.739	0.479	0.738	0.821
2.BINARY	460	0.827	0.896	0.732	0.857	0.931
3.ZSCALE	**115**	**0.812**	**0.985**	**0.810**	**0.887**	**0.957**
4.EGAAC	95	0.896	0.850	0.747	0.876	0.901
5.CTDD	195	0.996	0.077	0.184	0.682	0.655
6.CTDC	39	0.735	0.696	0.433	0.720	0.795
7.CTDT	39	0.823	0.712	0.541	0.779	0.827
8.CTriad	343	0.823	0.870	0.694	0.843	0.926
{1,3}	**135**	**0.896**	**0.908**	**0.806**	**0.902**	**0.958**
{2,3}	575	0.873	0.888	0.764	0.881	0.94
{3,4}	458	0.835	0.904	0.743	0.866	0.944
{3,8}	210	0.873	0.896	0.771	0.885	0.943
{3,5,6,7}	388	0.831	0.85	0.681	0.839	0.921
{2,3,4,8}	1013	0.85	0.908	0.762	0.876	0.947
{1,2,3,8}	938	0.742	0.985	0.751	0.844	0.938

Note: The first column represents the different feature extraction methods employed in this study. Dim refers to the different dimensions of every feature, and Sn, Sp, MCC, F1 and AUC represent the sensitivity, specificity, Mathews Correlation Coefficient and AUC value, respectively.

3.3 Z-scale feature analysis

In section 3.2, z-scale performed the best in the single feature comparison. The z-scale feature based on amino-acids' physical-chemical properties includes five z values. The distribution of amino acid residues around serine sites is able to determine the different physicochemical properties between phosphorylated sites and non-phosphorylated sites. From **Figure 2,** we can see that the z3 values of the phosphorylated sites are smaller than that of the non-phosphorylated sites, implying that a more negative charge occurred around viral protein phosphorylated sites than around non-phosphorylated sites. The results also showed that the z1, z2, z4, and z5 values of the phosphorylated sites are bigger than that of the non-phosphorylated sites. Overall, the different z-scale compositions surrounding the phosphorylated and non-phosphorylated sites indicates that it is reasonable to choose the z-scale as a feature for prediction.

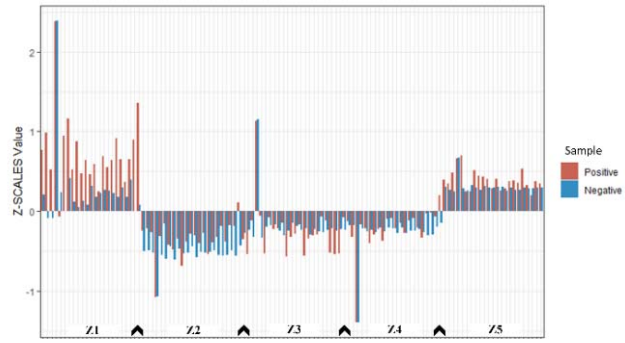

Figure 2. Comparison of the z-scale in positive and negative datasets. The vertical axis represents the z-scale values. The X-axis represents the five binary sequences.

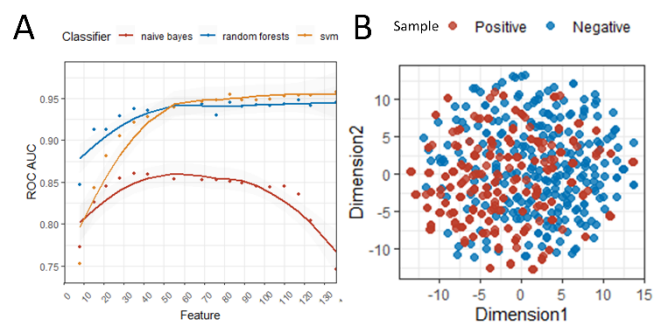

Figure 3. (A) Five-fold cross-validation performance of the three classifiers on different features. (B) t-SNE visualization of positive and negative data using 68D features.

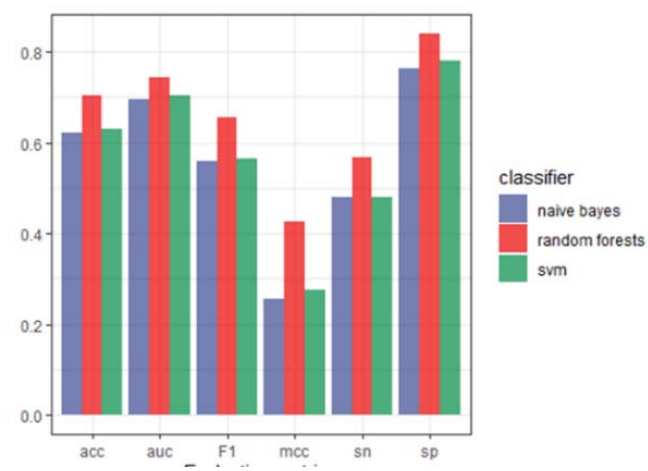

Figure 4. Independent test results. Sensitivity, specificity, AUC, MCC and F1-score of the proposed features in three classifiers.

3.4 Features selection and performance evaluation

VPTMdb provides three classifiers: support vector machine, random forest and naïve Bayes. Different dimensional features may have different impacts on different predictors. Thus, we selected features of different dimensions using the mRMR algorithm and compared the three

classifiers' performance from the 5-fold cross validation (**Supplementary Table S3**).

Figure 3A shows that the maximum AUCs of the svm and random forest are similar. For the random forest and svm, the AUCs increased when more features were selected (random forest: 14-135 features, with AUC > 0.90; svm: 27-135 features, with AUC > 0.90). However, we observed that the AUCs of naïve Bayes (AUCs > 0.80) decreased when more features were added. From a statistical point of view, to prevent the curse of dimensionality, fewer and more meaningful features should be chosen. Taking the above results into consideration, for 68 features, the AUCs of the three predictors perform better, suggesting that 68D is the most meaningful feature among all the features.

To understand the effective of our 68-dimensional features, the T-distributed Stochastic Neighbour Embedding (t-SNE) algorithm was used to visualize the positive and negative samples. A clear distinction was observed between the positive and negative samples, implying that our features selection results are effective (**Fig. 3B**).

To assess the robustness and performance of the svm, random forest and naïve Bayes in 68D features, 10-fold random independent tests were performed. The model performance on independent datasets is shown in **Figure 4**, random forest performed better, the average AUC, MCC, F1-score of its are 0.744, 0.427, 0.656 respectively. Comparing random forest and PSI-blast (**Supplementary Table S4**), the MCC, acc and sp values of random forest are higher than PSI-blast for 6.92%, 2.8% and 19.1%. Taking all indicators into consideration, our method is stable and better performance. We implemented svm, random forest and naïve Bayes into VPTMpre, users can choose them to predict phosphorylated sites of interest.

4 Conclusion

In this work, we constructed a conformable viral post-translational modification database, i.e., VPTMdb, which integrated 414 experimentally validated PTM sites from PubMed and 498 public PTM sites from the UniProtKB/Swiss-Prot database. VPTMdb is the first database that systematically collected viral protein PTMs. Virus-human PPI data were also collected in the VPTMdb to determine PTM sites association functions. These viral protein PTM data provide unique insights into virus-host interactions and are a valuable resource for biologists to recognize potential drug targets.

Moreover, VPTMpre, a novel feature representative classifier, was developed to predict viral protein serine sites. We compared various feature extraction methods and selected the optimized features using the mRMR algorithm. The feature analysis results showed that 68D was able to distinguish the phosphorylated sites and non-phosphorylated sites in viral proteins. VPTMpre was integrated into the VPTMdb web server to provide an online phosphorylation site prediction service. Users can choose three classifiers (svm, random forest and naïve Bayes) to predict phosphorylated sites of interests. However, because of data limitations, the prediction of VPTMpre is limited to serine sites. With continuous collection of new viral PTM data, we expect that VPTMpre will be extended to predict more types of PTM sites and obtain a better performance.

In the future, to respond to the rapid growth of viral PTM data, VPTMdb will be updated regularly and more viral PTM-related data collected to ensure that it provides the most comprehensive information to users. As the first attempt to develop the comprehensive viral PTM database, we sincerely welcome support and suggestions from the research community to improve the VPTMdb database.

Acknowledgment

Thanks for the anonymous reviewers for their kind suggestions.

Funding

This work has been supported by the Natural Science Foundation of China (No. 61771331, No. 61922020, No. 31572272), Youth Innovation Promotion Association, (Y62993) and the ten thousand plan - The national youth talent support program.

Conflict of Interest: none declared.

References

Ammari, M.G., *et al.* (2016) HPIDB 2.0: a curated database for host-pathogen interactions. *Database : the journal of biological databases and curation*, **2016**.

Banerjee, I., *et al.* (2014) Influenza A virus uses the aggresome processing machinery for host cell entry. *Science (New York, N.Y.)*, **346**, 473-477.

Chang, W., *et al.* (2018) shiny: Web Application Framework for R.

Chawla, N.V., *et al.* (2002) SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357.

Chen, Z., *et al.* (2018) iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics (Oxford, England)*, **34**, 2499-2502.

De Jay, N., *et al.* (2013) mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics (Oxford, England)*, **29**, 2365-2368.

Fu, L., *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, **28**, 3150-3152.

Govindan, G. and Nair, A.S. (2011) Composition, Transition and Distribution (CTD) — A dynamic feature for predictions based on hierarchical structure of cellular sorting. *Journal*, 1-6.

Hanchuan, P., *et al.* (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238.

He, W., *et al.* (2018) Research Progress in Protein Post-Translational Modification Site Prediction. *Briefings in Functional Genomics*, **18**, 220-229.

Hellberg, S., *et al.* (1987) Peptide quantitative structure-activity relationships, a multivariate approach. *Journal of medicinal chemistry*, **30**, 1126-1135.

Huang, G.H. and Li, J.C. (2018) Feature Extractions for Computationally Predicting Protein Post-Translational Modifications. *Current Bioinformatics*, **13**, 387-395.

Huang, K.Y., *et al.* (2013) ViralPhos: incorporating a recursively statistical method to predict phosphorylation sites on virus proteins. *BMC bioinformatics*, **14 Suppl 16**, S10.

Ivanov, A., *et al.* (2018) HIV-1 Tat phosphorylation on Ser-16 residue modulates HIV-1 transcription. *Retrovirology*, **15**, 39.

Krijthe, J.H. (2015) Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation.

Kuhn, M. (2020) caret: Classification and Regression Training.

Kulej, K., *et al.* (2017) Time-resolved Global and Chromatin Proteomics during Herpes Simplex Virus Type 1 (HSV-1) Infection. *Molecular & cellular proteomics : MCP*, **16**, S92-s107.

Liu, B. (2019) BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Briefings in bioinformatics*, **20**, 1280-1294.

Randow, F. and Lehner, P.J. (2009) Viral avoidance and exploitation of the ubiquitin system. *Nature cell biology*, **11**, 527-534.

Scaturro, P., *et al.* (2018) An orthogonal proteomic survey uncovers novel Zika virus host factors. *Nature*, **561**, 253-257.

Schwartz, D. and Church, G.M. (2010) Collection and motif-based prediction of phosphorylation sites in human viruses. *Science signaling*, **3**, rs2.

Shen, J., *et al.* (2007) Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 4337-4341.

Sing, T., *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)*, **21**, 3940-3941.

van der Maaten, L.J.P. and Hinton, G.E. (2008) Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579-2605.

Wickham, H. (2016) ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*.

Zheng, J., *et al.* (2018) Identification of N-linked glycosylation sites in the spike protein and their functional impact on the replication and infectivity of coronavirus infectious bronchitis virus in cell culture. *Virology*, **513**, 65-74.